DOCUMENT RESUME

ED 349 732 EC 301 477

AUTHOR Harnisch, Delwyn L.; And Others
TITLE Human Judgment and the Logic of Evidence: A Critical Examination of Research Methods in Special Education Transition Literature.
PUB DATE 92
NOTE 27p.; In: Harnisch, Delwyn L., And Others. Selected Readings in Transition; see EC 301 473.
PUB TYPE Viewpoints (Opinion/Position Papers, Essays, etc.) (120)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Causal Models; *Disabilities; Education Work Relationship; *Evaluative Thinking; Experimental Groups; High Schools; Influences; *Quasiexperimental Design; Research Design; *Research Methodology; Transitional Programs; *Validity

ABSTRACT

This paper describes several common types of research studies in special education transition literature and the threats to their validity. It then describes how the evidential base may be broadened, how diverse sources of evidence can be combined to strengthen causal inferences, and the role of judgment within quasi-experimentation. The paper discusses issues internal to studies and to the methods used in conducting the studies, and discusses issues arising when attempts are made to use the results of the study with other groups or in other places. Threats to internal and external validity are examined. True experimental design is outlined, and then types of quasi-experimental designs are described, including one-group posttest-only design, one-group pretest-posttest design, comparison-group pretest-posttest design, prematched control group design, natural experiments, longitudinal research, cross-sectional research, case-study and single-subject designs, and meta-analysis. Technical and conceptual advances that provide a more significant basis for the interpretation and limitations of quasi-experimental designs are explored. Three considerations required in developing a compelling argument about the causal influence of an intervention are discussed: the analysis must provide a well-specified and credible rationale that links the causal mechanisms with outcomes; it must present evidence to substantiate the claim that the intervention is a plausible explanation for the observed outcome; and it must provide diagnostic assessments and establish the value of the information about purported causal mechanisms and rival explanations. (Contains approximately 30 references.) (JDD)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

ED349732

Chapter 4

# Human Judgment and the Logic of Evidence: A Critical Examination of Research Methods in Special Education Transition Literature

Delwyn L. Harnisch

Adrian T. Fisher

and

Michael L. Connell

RUNNING HEAD: Research Methods

EC 301477

BEST COPY AVAILABLE

**Human Judgment and the Logic of Evidence: A Critical Examination of Research Methods in Special Education Transition Literature**

Correct reading and interpretation of research articles have significant implications for both applied and basic research areas, because the results of such research can guide many forms of decision making. Policymakers, for example, must build up a basis of research that can explore the theories to explain actions, as well as to define the nature of problems that have been identified. In contrast to a single experiment or case study, a well-conducted program of research contributes much to such a knowledge base.

Teachers and those who work in educational and other service agencies often use the results of applied studies to help them develop and implement new curricula and service delivery methods. These people must understand the potential strengths and weaknesses of such studies and establish whether extraneous factors were ruled out to ensure that results are applicable to the reader's clients and settings. Based on such understanding, potential applications of the technology, services, type of interventions, and outcomes of a given study can be examined systematically.

The first part of this paper describes several common types of research studies and the threats to their validity. We will then describe how the evidential base may be broadened, how diverse sources of evidence can be combined to strengthen causal inferences, and the role of judgment within quasi-experimentation.

## Common Threats in Empirical Studies

When evaluating research studies, two major factors should be considered: internal and external (or generalizability) validity. First, those issues internal to the study and the methods used in conducting the study will be discussed; second, those issues arising when attempting to use the results of the study with other groups, or in other places. The various factors that can lessen either of these in a study are called threats to validity.

In examining the internal validity of an experiment, we are attempting to find out whether the results occurred because of the intervention used or because of some unconsidered factors occurring during the experiment. In this connection, one should always consider how events other than the planned intervention might affect the results of the experiment. For example, subjects' normal development (e.g., growing older, stronger) may not be a part of the formal experiment, but can affect the results.

External validity, in turn, refers to determining the extent to which an experiment's conclusions are transferrable to other subjects and settings. Determining the external validity of a study is crucial when using research findings to develop educational programs, approaches, or curriculum materials. If the study setting or the subjects are not comparable, the results may not transfer to other situations.

Threats to internal and external validity are always present and must be controlled in every study. A well-designed study, therefore, attempts to achieve the optimum balance between what one would ideally need to do to control these threats and what one can actually do in a real-world setting or in situations where analyses are limited to secondary analysis of extant databases. However, as real-world events often dictate the way studies are actually conducted, it is important to remember that the research designs reflect various compromises between and among these issues.

Every study is conducted according to a plan or research design that specifies the information to be gathered and how it will be gathered, as well as selection and assignment of subjects to groups. Usually, one would first construct a hypothesis to investigate. A hypothesis can be a simple statement in which we predict that if we do one thing (e.g., introduce a new curriculum), then there will be a certain outcome for the subjects (e.g., an improvement in test scores). This statement can be described as "If X, then Y"; in this statement, X is the intervention, treatment, or change and is termed the independent variable, while Y is the outcome or dependent variable.

After an observable hypothesis has been formulated, the experimental conditions (independent variables) are specified together with any identifiable nuisance factors (threats to validity) that might interfere with the experiment. With these items identified, the number of subjects, the populations from which the subjects are taken, and the assignment of selected subjects into experimental groups are determined. Finally, the measurements to be recorded for each subject and the analyses to be performed are specified.

Internal Validity

The researcher selects a research design to control as many of the identified nuisance variables as possible, while increasing the power of the measurements being made. In selecting a design, the manner in which subjects are selected or assigned to groups should be carefully examined. Likewise, the subjects' experiences after assignment to treatments should be examined. For example, consider the impact that newly mandated programs might have upon already existing experiments. A federally mandated curriculum requiring that high school students with learning disabilities use word processors could easily invalidate existing investigations of the effectiveness of technology use for this population. Observed differences in outcome measures may reflect pre-existing differences between the subjects assigned to the various experimental treatments rather than experimental effects. Random assignment to groups, for example, is characteristic of "true" experimental designs, as it controls many threats to internal validity by eliminating selection bias.

The environment in which the experiment is conducted also plays an important role in determining internal validity. In some studies, information may be selectively presented to subjects in the experimental group, but not to others. If the subjects in the different groups are able to communicate with one another, they may share that information, thereby possibly contaminating experimental differences. Actions of people outside the experiment can also serve to damage internal validity. For example, a teacher who is not part of the intervention portion of the experiment may provide special tutoring to students in a control group who are observed to be falling behind those in the experimental program.

An illustration of how this may be examined was provided by Edgar in a 1988 grant proposal submitted to OSERS. In this proposed study, 100 students identified as potentially at risk were randomly assigned to one of three conditions: a case management system, a mentor program, or a peer social group. Environmental effects were monitored by collecting a wide variety of longitudinal post-school variables from interviews with a representative sample from the surrounding area and the population of special education students.

Edgar described the value of using multiple data points to increase the accuracy of the student profile over time. Some of the additional data points included rate of chronic absenteeism, number and nature of suspensions, and number and type of academic credits earned. To further capture the influence of environmental factors on the subjects, Edgar proposed that the data-collection phase would begin while the subjects were in the early years of their scholastic career.

Other threats to internal validity are directly related to the methods used in the experiment. For example, reliance upon repeated testing using the same type of evaluation often results in subjects learning the test rather than the subject matter which, in turn, affects the experimental outcome. Observed findings may also reflect addition or loss of subjects during the experiment rather than the experiment itself. A common statistically based threat to internal validity is regression. Regression, in this sense, is used to describe the statistical drift of extreme scores towards the group average score; that is, individuals who have extreme scores at first testing statistically tend to achieve scores that are less extreme upon subsequent testing.

External Validity

A common threat to external validity is posed by improper selection of the outcome measure, the dependent variable. The outcome measures must be carefully selected to ensure the closest match between theory and practice. Another obstacle to external validity relates to determining the population to be used and the extent to which results from that

population can be generalized. This part of the experiment can be broken down in at least three points: defining the population to be used, identifying and locating its members, and sampling adequately from the members who have been located. If the population cannot be defined, adequately sampled, or located, the results of a given study apply only to the cooperating sample studied.

An important extension of external validity is ecological validity. This concern must be addressed if the findings of an experimental study are to be applied to real-life settings. Often experimental findings are based upon artificial samples of places, times, or social demands. When the participants or results of such studies are transferred to real life settings, the experiences break down.

Studies of the long-term effects of social skill training on placement in deinstitutionalized settings serve as prime examples of the need to be aware of ecological validity. Keith, Schalock, and Hoffman (1986) and Schalock (1986) reported that the subjects found to have the most successful transition, as measured by the need to be returned to a more restrictive living environment, were those coming from the transition training programs that most closely approximated community life. Conversely, those subjects who received social skill training in a restrictive environment that did not approximate the experiences and demands of actual real-life settings experienced less successful transitions to community living and, therefore, were more likely to be returned to institutions.

The myriad threats to validity and reliability must be balanced against the practicality of conducting studies in the real world by using the various ways in which research can be conducted. Researchers may select from a broad number of experimental designs. Some of these designs rely on a restricted range of subjects and measures, whereas others can take advantage of the tight validity control that is inherent in the so-called "true experimental designs."

## True Experimental Design

True experimental designs make use of a comparison technique in which at least one group of subjects, the experimental group, receives treatment (e.g., a new training curriculum). At least one other group, the control group, continues to receive the normal treatment (e.g., the current curriculum). Each group of subjects is measured on the outcome variables before the intervention to provide a baseline for comparison and again after the intervention to assess any changes that may have occurred. Many other research designs utilize control groups to help assess the impact of treatment, but true experiments are characterized by random assignment of subjects to the various experimental and control groups.

In the simplest type of experimental design, often referred to as a completely randomized design, the subjects constituting the experimental and control groups have an equal chance of being assigned to an experimental and a control group. That is, the random assignment is subject to no restriction other than the option of assigning each treatment level to the same number of subjects.

A slightly more complicated design uses an initial blocking procedure to deal with an identified nuisance variable. In many research studies, the subjects have markedly differing attitudes, experiences, and abilities that may have an impact upon the results of the intervention. Although such differences are often present, they may be dealt with through appropriate blocking of groups. For example, in examining the effectiveness of a proposed curriculum, it might be helpful to place all subjects with high test scores in one subgroup, those with medium scores in another, and those with low scores in a third. Once the blocks are created, the subjects within each are randomly assigned to experimental or control groups.

Other commonly used blocking procedures include the use of gender, disabling condition, and severity of disability. Designs using this approach are referred to as completely randomized block designs.

Many other true experimental designs are available, each offering a slightly different set of advantages and trade-offs. Authors should identify clearly the experimental design they have used to allow readers to utilize such information in their evaluation of the research. For example, Heal, Colson, and Gross (1984) thoroughly discussed both the experimental design and the subsequent analysis in their study of training effects for students with severe mental retardation.

## Quasi-Experimental Designs

The majority of transition studies use a quasi-experimental design. This poses more significant compromises between validity controls and practicality compared to the more formal true experimental designs. Thus, research comparing alternative treatments provides a stronger basis for inference about the effects of a given intervention if it is conducted using true experimental designs based on random assignment to groups.

A wide variety of research designs fall under the quasi-experimental heading. Some closely resemble true experimental designs, except that they do not use random assignment of subjects to groups; others use only one group and limited testing.

### One-Group Posttest-Only Design

In studies using the one-group posttest-only design, the planned intervention is performed and an outcome measurement is collected and analyzed. There is no measure of the level of achievement before an intervention, nor is there any group against which to compare the results for the subjects in the experimental group.

Dalke and Schmitt (1987) used this approach in their study of academic preparedness and college skill transition training methods for students with learning disabilities. After participating in a special summer program, the subjects were asked to complete a 17-point questionnaire and their student diagnostic profiles were re-evaluated.

Like the other quasi-experimental designs, this design cannot confirm causal relationships. No pretests are given and, as a result, no comparisons are made with control or other groups receiving alternative treatments. The absence of a comparison between

groups and the lack of baseline data are fundamental weaknesses of this design. Without such comparisons, we cannot be certain that the intervention caused the changes—or that, indeed, there was any change at all.

An additional significant weakness of this design is the lack of control for selection biases. As a result, the design is most appropriate only for simple descriptive studies, as there are no satisfactory controls for threats to internal validity, especially selection biases.

One-Group Pretest-Posttest Design

In studies using one-group pretest-posttest design, subject performance is measured on the dependent variable before the intervention. After the intervention is conducted, performance is again measured on the dependent variable. As with the one-group posttest-only method, there is no group against which to compare intervention results.

The addition of a pretest assessment provides an improvement over the posttest-only design. Comparisons of changes in the assessment results allow evaluation of changes in the dependent variable. However, statements regarding treatment effect cannot be supported. Although the pretest measurement provides a baseline of performance making it possible to detect change, threats to internal validity are not adequately controlled by the use of a single group.

Comparison-Group Pretest-Posttest Design

Addition of pretest measures strengthens internal validity by partially controlling some extraneous variables. Inclusion of both a pretest and a comparison group can increase interpretability of treatment effects, even when no attempt is made to make the members of the two groups comparable on many salient variables.

This method was followed by Collins, Engen-Wedin, Margolis, and Price (1987), who used data from three sections of a writing class using word processors. The classes from which the subjects were drawn contained 22 students with learning disabilities and 52 without, forming the basis for creating two groups for comparison. In their analysis, the

authors assessed writing assignments from before and after the intervention to measure the outcomes.

This design allows group posttest differences to be compared more readily. To some extent, we are able to use these data to evaluate how effective the intervention itself has been. However, because this type of study does not use a control group that matches the experimental group, it is difficult to determine whether the findings resulted from the intervention or from factors more related to the subjects themselves.

Prematched Control Group Design

In this design, treatment and control groups are matched after pretest evaluations and the intervention is implemented. Such matching may be performed on the basis of disabling conditions, test results, or other common elements. Treatment effects are assessed by comparing posttest scores, or the change in scores between the pretest and posttest for each group.

Two flaws are especially threatening to prematched control-group designs: selection interactions and statistical regression. Although the groups had been assessed and matched for equivalence, one cannot assume that the entire array of relevant variables were held constant. For example, posttest differences could be explained by interactions of such factors as maturation and history. However, the greatest threat to the validity of the matched-group design is statistical regression. Regression here describes the statistical drift of extreme scores toward the population average; that is, statistically, individuals who have extreme scores at the first testing tend to obtain scores that are less extreme upon subsequent testing.

Natural Experiments

Studies using naturalistic designs are typically used to test hypotheses. A number of questions are asked, and descriptive analyses are completed in an effort to discover associations among variables. The associations are interpreted, and hypotheses of causation are proposed.

Goldberg's (1986) study of coping strategies used by students with learning disabilities provides a good example of this style of inquiry. This exploratory study used a wide variety of psychoeducational assessments, interview data, and examination of work products to provide descriptive data about students with learning disabilities.

A related research technique, often described as naturalistic, has been developed in the fields of ethnology and anthropology. This technique involves observing people in their natural environment as unobtrusively as possible. It differs from the surveys and direct observations described here as the measures used are often developed as part of the observation procedure.

Longitudinal

Longitudinal research might be viewed as a form of the one-group pretest-posttest design. The first step in this method involves measuring subject performance on an outcome measure. After establishing this baseline, an actual intervention is offered or time is allowed for natural development to occur, or both. At the end of a specified period, group subject performance is again determined for the outcome measurements. Use of this type of study is important to the understanding of the long-term impact of interventions on those who received them. The design suffers from the threats to internal validity outlined in the discussion of the one-group pretest-posttest design.

Bireley and Manley (1980) used the longitudinal approach in their investigations of 10 students in Wright State University's program for individuals with learning disabilities over the first two years of the program. The outcome measures consisted of rates of retention, grade point average, and numbers leaving the university.

Cross-Sectional

In cross-sectional research, a sample is drawn from the population of interest and selected outcome measurements are obtained. After the passage of time a second sample (not necessarily consisting of the same members as the first sample) is drawn, and the desired measurements are again taken.

An example of this research technique is found in Allen (1986), who analyzed the data on the performance of students with hearing impairments collected across the United States during two major norming studies. Although the two groups did not include the same subjects, his analyses of these data provide helpful information about the relative performances of these groups of students over a 10-year period.

Case-Study and Single-Subject Designs

Case-study and single-subject designs consist of an intense, detailed description and analysis of a single individual, project, program, or instructional material in the context of its environment. By nature, these designs control most threats to internal validity. Specifically, selection bias is perfectly controlled as the experimental and control conditions are present in the same subject. History is controlled by repeating intervention and baseline alterations or by varying the time at which intervention begins in different areas. Maturation is assessed by ongoing measurements; intervention effects can be seen against the baseline of growth or degeneration, if any. Finally, regression effects are controlled by extending baseline measurements until they become stabilized about their "true score" values.

However, measurement bias and reactivity form serious threats to these designs. Especially problematic is the repeated measure by experimenters who know their subjects extremely well.

The major drawback of case-study and single-subject design studies is the threat to external validity. Because only one subject is examined at a time, there is no way to equate the results to others. This limits the use of the described procedures in dealing with other subjects. Another threat to internal validity is the Hawthorne effect, whereby observed changes result from subjects' attempts to respond to the experimenter rather than from the interventions.

Meta-Analysis

The strategy combines the results of all studies that have tested essentially the same hypothesis. Meta-analysis can be conducted statistically by converting the reported statistics to a common metric for re-analysis, or in the more common form of an extensive, critical literature review.

A serious disadvantage of this technique is the difficulty in maintaining internal and external validity. For external validity, the meta-analysis must identify the population of studies that have tested a particular hypothesis. Published studies are almost certainly biased in favor of those in which a significant effect was found, and the extent of this bias cannot be estimated. For internal validity, the meta-analysis must combine the results from studies whose procedures and statistical approaches varied greatly from one another.

Despite potential problems when viewed as a scientific method, meta-analysis is a more objective and public procedure than the integrative literature review. It can result in valuable synthesis of information, as is evidenced in the work of Cook, Scruggs, Mastropieri, and Casto (1986), who conducted a meta-analysis of available research documenting the effectiveness of using students with disabilities as the tutors of other students. Implications for instruction and further research from this analysis were provided.

## Information-Gathering Techniques

Regardless of whether a study employs a true or a quasi-experimental design, many techniques and methods are available for gathering the outcome measurements. A number of studies will be described here as examples of effective use of data-gathering tools that may be used in a wide variety of inquiry.

For example, Salend and Fradd (1986) provide an excellent example of the use of survey data in an educational study. These researchers gathered data through a survey questionnaire and follow-up telephone calls to the Commissioners of Education in each of the 50 states and the District of Columbia. Despite the difficulty inherent in getting a high

participation rate in surveys, Salend and Fradd's results were based on 50 of the 51 Commissioners contacted.

Another use of survey data may combine a number of the methods described above. Fisher and Harnisch, in the current volume, used data from the High School and Beyond survey in a longitudinal study to examine over time (a) the career aspirations of youth with and without disabilities, (b) the differences between the two groups, and (c) the changes that occurred over time. This study is a pretest-posttest, nonequivalent comparison group study.

Survey data such as those of High School and Beyond may also be used to test theories and develop hypotheses that can later be used in applied settings to design programs or components. Principal-components analysis, factor analysis, and their many variations are often used to allow the investigator to determine which of a large number of variables cluster together to form a much smaller number of dimensions. Once these clusters are known, they provide target areas for developing applied strategies and further research questions.

Other types of archival data, such as grades, medical records, and case histories, are important and, therefore, often utilized. For example, Friedrich, Fuller, and Davis (1984) used approximately 1,600 student referrals to investigate the discriminating power of 96 empirically derived formulas for assessing learning disability. Data of this type can help provide a more general discussion of subjects and provide the basis for constructing groups for discussion purposes.

## Expanding the Evidential Base

So far, the discussion has dealt with various ways in which empirical research studies can be conducted, yet the major approach throughout has been quasi-experimentation. Consideration of threats to validity aids in interpretating and using information from such studies, but technical and conceptual advances in quasi-experimentation now provide a more significant basis for the interpretation and limitations of these designs.

Methods of statistical analysis have become increasingly sophisticated, allowing us to estimate parameters in complex cause-and-effect models. Moreover, improved diagnostic tests enable us to better determine if (and how well) data fit these models. To further offset the other imperfections in quasi-experimental analysis of causal relations, the use of multiple strategies (e.g., methods, measures, analysts) has been widely advocated.

Despite a continued series of advances, evaluations following the quasi-experimental paradigm still exhibit serious flaws. Although it is reasonable to expect that some fraction of studies will be inadequate, the transition literature appears to contain a disproportionate number of poor studies. Few of the studies reviewed are relevant, credible, and reported well enough to be used for examining policy issues concerned with the effects of specific intervention programs.

Reported weaknesses are not isolated to particular substantive areas (Gilbert, Light, & Mosteller, 1975; Lipsey, Crosse, Dunkle, Pollard, & Stobart, 1985). They have been reported in assessments of youth employment training programs (Betsey, Hollister, & Papageorgious, 1985), education (Boruch & Cordray, 1980), maternal and child health (Shadish & Reis, 1984), and juvenile justice (Maltz, Gordon, McDowall, & McCleary, 1980). The relatively high incidence of technically poor studies poses a serious threat to the reputation of the field.

What factors have contributed to this state of affairs? Some programs may not have been well enough developed to enable meaningful experimentation. Studies included in the reviews may have been planned and conducted long before sophisticated technology was available. Perhaps we are expecting too much of social-science methods; that is, they may inherently be too crude to match the complexity of social programs. Or, as a profession, perhaps we simply have not learned when and how to conduct these assessments properly. Each of these reasons contributes to understanding the problem better while implying a different set of solutions.

The effects of intervention were evaluated initially with an experiment in which the effects of the intervention were assessed. Evaluations followed this perspective (input-output assessment), in large part because of the conceptual simplicity of the process of developing and summarizing information. Such evaluation plans were relatively simple, consisting primarily of: (a) selecting suitable measures, (b) devising an assignment plan, and (c) managing the implementation of these key features.

Using this model, inference about program effects stemmed from tests of statistical significance applied to data derived from randomized experiments. The development and synthesis of evidence about program effectiveness using the experimental paradigm implicitly mixes these two processes, thus removing the need for judgment on the part of the researcher.

Despite forceful warnings of inferential weaknesses (Campbell & Boruch, 1975; Cook & Campbell, 1979), quasi-experiments have been treated merely as impoverished versions of true experiments, the chief difference between the two being the lack of random allocation to conditions. In contrast to the probing, searching, active testing of the plausible effects of rival explanations described by Campbell and his co-workers (Campbell, 1969, 1984; Campbell & Stanley, 1966; Cook & Campbell, 1979), early studies seemed to focus on attempts to find approximate statistical models to control for influence of pretreatment differences. Kenny (1975) pointed out that chance is only one rival explanation.

Two problems are obvious from such analyses. First, early evaluations using quasi-experiments were based on a limited notion of what constitutes evidence about a program's effectiveness. Thus, evidence of program effectiveness was limited largely to establishing one fact: Did the treatment group outperform the control group?

A test of statistical significance was usually presented in support of a claim. However, several intermediate facts must be established before a causal claim can be justified. For example, were the conditions necessary for change present? Was the appropriate clientele exposed to the intervention? Was the intervention properly implemented? Was the

intervention implemented with sufficient intensity to trigger the causal chain of events necessary to induce a change in behavior? Each of these questions requires that we decompose the treatment package into its elements. Judging from the reviews of the literature (Harnisch, Chaplin, Fisher, & Tu, 1986; Harnisch, Fisher, Kacmarek, & DeStefano, 1987; Lipsey et al., 1985), explorations of the "black box" of program treatments are relatively rare.

Second, current quasi-experimental analysis assumes a passive posture toward development and synthesis of evidence about causal claims. This posture is manifest in three widespread beliefs: (a) that nonequivalent group designs can and do control for threats to validity; (b) that statistical procedures (for example, tests of significance, adjustments for nonequivalence) perform as intended; and (c) that assumptions are robust enough to be safely ignored. To augment this analysis, one rarely sees discussion on the adequacy of the statistical design for an evaluation, while the assumptions are often stated as caveats rather than being probed with additional design elements.

Judgment Within Quasi-Experimentation

A review of the empirical literature suggests that the role of judgment within quasi-experimentation has neither been fully acknowledged nor properly employed in practice. Herein lies one of the fundamental problems in current quasi-experimental analysis.

When evaluating the logic of evidence used to test cause-effect relationships, it is generally believed that causal relationships are established if three conditions hold. First, the purported cause (X) precedes the effect (Y); second, X covaries with Y; third, all other rival explanations are implausible. An ideal case where all three conditions are met allows us to state a fact (the treatment caused an increase in performance) with the separate effects of artifacts held in check. The third condition plays an especially important role in causal inference. The credibility of the evidence about a causal claim is greatest when no plausible alternative explanation can be invoked; it is lower when such alternatives are available.

Causal inferences derived from quasi-experimental analyses rarely satisfy this condition; that is, the internal validity of the inference is always suspect.

Our view of causal evidence is inherently limited if the most distinctive feature of causal analysis is the need to discount the influence of other factors. Einhorn and Hogarth (1986) likened the diagnostic value of discounting other explanations to the case of the mystery writer who reveals only who did not commit the crime. Similarly, covariation of cause and effect is too simplistic a criterion when X is part of a complex set of factors that influence Y. And, although X must occur before Y occurs, temporal contiguity is low or ambiguous in many field applications. Therefore, although the classic criteria for establishing causal relationships may be adequate guides for developing evidence in relatively closed systems, a more comprehensive set of guiding principles is needed for quasi-experimental analysis in open systems like program research on disabled populations.

If we grant that the criteria for establishing causal relationships are impoverished, the question then becomes: On what grounds can we derive a more comprehensive notion of evidence within quasi-experimental assessment? One way to approach this question is to look at the nature of the judgmental tasks that an analyst must perform. In practice, quasi-experimental analysis falls somewhere between pure reliance on scientific methods and pure human judgment. A reasonable set of principles regarding evidence within quasi-experimentation must take this mixture of methodology and judgment into account. In particular, issues about evidence appear in two distinct tasks: the development of a data-acquisition plan and the synthesis or combination of evidence into a coherent set of results. In both tasks, the analyst exerts considerable discretion over the evidence to be included, its completeness and relevance, and how it should be combined and presented in making a summary judgment about the strength of the causal relationship.

The analyst is often required to derive conclusions about the effects of an intervention by piecing together numerous bits of information accumulated by multiple methods—a process akin to Sherlock Holmes' investigative tactics (Larson & Kaplan, 1981; Leamer,

1978). Because many issues implied by these practices fall outside the domain of classical statistical theory, proposed solutions to these combinatorial procedures have been sparse.

Researchers grappling with these issues (Fennessey, 1976; Finney, 1974; Gilbert, Mosteller, & Tukey, 1976) have identified many problems faced by users of multimethod strategies, for example, nonindependence of evidence and the resulting overconfidence in conclusions, judgments about the differential credibility of evidence, and data-instigated specification searches. The questions then becomes: How can complex and diverse sources of evidence be combined to form an overall judgment of the strength of a causal relationship for a transition program? Are some intuitively appealing transition procedures subject to inferential difficulties? The answers to these questions depend on the types of methodologies employed and the degree to which human judgment is involved.

We begin by examining the systematic rules that people use in judging ordinary causal relations. Judgment plays a central role in quasi-experimental analysis. For example, an examination of the evidence on stereotypical biases or flaws that individuals exhibit can lead to corrective solutions on the development and synthesis tasks. The results of an analysis in applied research are often intended to be used by others, such as policymakers. Having an understanding of the way in which causal evidence is interpreted can also help ensure that the evidence developed is maximally credible and useful.

Einhorn and Hogarth's (1986) review of the literature on judging probable cause asserted that scientific and ordinary causal inferences are made within the context of both a causal field and existing interrelationships among several cues-to-causality (that is, temporal order, distinctiveness, strength of the causal chain, covariation, congruity, and contiguity). When these factors are combined, they determine one's perception of the overall gross strength of the causal relation.

Einhorn and Hogarth's formulation of the psychology of judging probable cause has several important implications for the ways in which we conduct and disclose formal causal assessments of the effects of interventions. First, the relevance of a particular causal

explanation (the treatment of rival explanation) depends critically on its role within a causal field, that is, on a specified set of contextual factors. The causal field sets the context for interpretation of difference among variables and deviations from expectation or steady states, and limits or expands the number and salience of alternative explanations. For a cause to be plausible, its distinctiveness from the background must be considered within the particular causal field. For program research with special populations, this means that the strength and fidelity of the treatment (relative to no-treatment conditions) must be determined. However, this is rarely done in practice (Scheirer & Rezmovic, 1983).

The Einhorn and Hogarth (1986) model also suggests that covariation need not be perfect in order to instigate a causal inference. They express the complex scenario where X is a necessary but not a sufficient part of the complex scenario that is itself unnecessary but sufficient to produce Y; this means that other causes of Y exist and only a specific set of conditions conjoins with X to produce Y in a given causal field. What these conditions are in practice depends on the program model, the theory, and the particulars of the setting.

The criteria of this model differ from the classical criteria in their explicit recognition of the need to establish causal chains to account for the overall strength of a relationship. Within this notion are the interdependent factors, contiguity and congruity. Contiguity refers to the extent to which events are contiguous in time and space. When contiguity is low (for example, when substantial time elapses between the presence of X and the appearance of Y), a causal relation is difficult to justify unless intermediate causal models are established to link the events. Congruity refers to the similarity of the strength (or duration) of cause and effect. In its simplest form, the notion of congruity implies that strong causes produce strong effects and that weak causes produce weak effects. This explanation, of course, is too simple. To account for seemingly anomalous relations (for example, small causes that produce big effects), additional processes must be specified that justify how the cause must be amplified (large effect, given a small cause) or dampened (small effect, given a large cause) to produce the observed magnitude of effect.

When considered together, contiguity and congruity form the basis for specifying the length of the causal chain necessary to link X with Y. When both are high, few if any links are needed. When congruity is low and contiguity is high, the mechanisms that dampen or amplify the effect must be considered. In the reverse case, links that bridge the contiguity gap are necessary. The most complex case is that in which both contiguity and congruity are low. Here, intermediate causal links are needed both to bridge the temporal gap and to represent the amplification or the dampening process.

Implications for Quasi-Experimental Designs in Transition Studies

The psychology of judging probable cause makes it clear that the types of evidence brought to bear in causal analysis cannot be limited to the simplistic input-output conception suggested by the three classic cues to causality discussed earlier. This is particularly true for quasi-experimental analysis, which usually does not rule out all rival explanations. To the extent that policymakers can muster their own rival explanation or that the findings are uncertain, the credibility of the results can be questioned or, worse, the findings can be disregarded entirely. For example, in the absence of sufficient detail about the transition process, it is legitimate to ask: How did this small treatment, installed in a "noisy" environment, cause a harmful effect on performance? One obvious answer—right or wrong—is that there must be something wrong with the methods used to derive the inference. Indeed, if a plausible model cannot be postulated, this seems to be a reasonable answer.

To evaluate the effect of a program that shows no treatment effect, we must have evidence that the treatment (that is, the cause) was indeed present and that the methodology was sensitive enough to detect any effect it may have produced.

Summary

Research findings can and should play a large role in guiding and directing decision making. The influence of research findings may be felt from creation of policy to implementation of curricular change and service delivery methods. These are appropriate

applications of research, yet findings are often utilized without concern for elements that might temper their application. Potential threats to a study's validity—both internal and external—should always be examined before one attempts to apply the results of a given study. Similarly, results from a single study should not be accorded the same weight and consideration as a systematic, well-conducted program of research.

In examining the transition literature, the notion of evidence within quasi-experimental analysis should be extended beyond the prevalent cues to causality established within the classic experimental paradigm. The comments in the preceding section suggest that a comprehensive view of evidence within quasi-experimental analysis requires at least three additional considerations to develop a compelling argument about the causal influence of an intervention. The analysis must first provide a well-specified and credible rationale that links the causal mechanisms with outcomes; second, it must present evidence to substantiate the claim that the purported causal agent (the intervention) is, itself, a plausible explanation for the observed outcome; and third, it must provide diagnostic assessments and establish the value of the information about purported causal mechanisms and rival explanations. In other words, we have to substantiate the basis for our conclusions through additional forms of evidence.

References

Allen, T. E. (1986). Patterns of academic achievement among hearing impaired students: 1974-1983. In A. Schildroth & M. A. Karchmer (Eds.), Deaf children in America (pp. 161-206). Boston: Little, Brown & Co.

Betsey, C. L., Hollister, R. G., Jr., & Papageorgious, M. R. (Eds.). (1985). Youth employment and training programs: The YEDPA years. Washington, DC: National Academic Press.

Bireley, M., & Manley, E. (1980). The learning disabled student in a college environment: A report of Wright State University's program. Journal of Learning Disabilities, 13, 12-15.

Boruch, R. F., & Cordray, D. S. (Eds.). (1980). An appraisal of educational program evaluations: Federal state and local agencies. Washington, DC: U.S. Department of Education.

Campbell, D. T. (1969). Prospective: Artifact and control. In R. Rosenthal & R. L. Rosnow (Eds.), Artifacts in behavioral research (pp. 351-378). New York: Appleton-Century Crofts.

Campbell, D. T. (1984). Can we be scientific in applied social science? In R. F. Conner, D. Altaman, & C. Jackson (Eds.), Evaluation studies review annual (Vol. 9, pp. 195-296). Beverly Hills: Sage.

Campbell, D. T., & Boruch, R. F. (1975). Making the case for randomized assignment to treatment by considering the alternatives: Six ways in which quasi-experimental evaluations in compensatory education tends to underestimate effects. In C. A. Bennett & A. A. Lumsdaine (Eds.), Evaluation and experiment: Some critical issues in assessing social programs (pp. 195-296). New York: Academic Press.

Campbell, D. T., & Stanley, J. C. (1966). Experimental and quasi-experimental designs for research. Chicago: Rand McNally.

Collins, T., Engen-Wedin, N., Margolis, W., & Price, L. (1987). Learning disabled writers and word processing: Performance and attitude gains. Research and Teaching in Developmental Education, 4, 13-20.

Cook, S. B., Scruggs, T. E., Mastropieri, M. A., & Casto, G. C. (1985-86). Disabled students as tutors. Journal of Special Education, 19, 483-492.

Cook, T. D., & Campbell, D. T. (1979). Quasi-experimentation: Design and analysis issues for field settings. Chicago: Rand McNally.

Dalke, C., & Schmitt, S. (1987). Meeting the transition needs of college-bound students with learning disabilities. Journal of Learning Disabilities, 20, 176-180.

Edgar, E. (1988). A polymorphic tracking and intervention model for students who drop out or are at risk of dropping out of special education programs in suburban Washington state. (Grant proposal submitted to OSERS, Washington, DC). Seattle: University of Washington, Experimental Education Unit WJ-10.

Einhorn, H. J., & Hogarth, R. M. (1986). Judging probable cause. Psychological Bulletin, 99, 3-19.

Fennessey, J. (1976). Social policy research and Bayesian inference. In C. C. Abt (Ed.), The evaluation of social programs (pp. 269-282). Beverly Hills, CA: Sage.

Finney, D. J. (1974). Problems, data and inference. Journal of the Royal Statistical Society, Series A, 137, 1-23.

Gilbert, J. P., Light, R. J., & Mosteller, F. (1975). Assessing social innovation: An empirical base for policy. In C. A. Bennet & A. A. Lumsdaine (Eds.), Evaluation and the experiment: Some critical issues in assessing social programs (pp. 39-194). New York: Academic Press.

Gilbert, J. P., Mosteller, F., & Tukey, J. W. (1976). Steady social progress requires quantitative evaluation to be searching. In C. C. Abt (Ed.), The evaluation of social programs (pp. 295-312). Beverly Hills: Sage.

Goldberg, R. (1986). The learning disabled college student: Problem areas and coping strategies. AHSSPE Bulletin (Vol. 1-4, 1983-1986). (ERIC Document Reproduction Service No. ED339335.)

Harnisch, D. L., Chaplin, C. C., Fisher, A. T., & Tu, J. J. (1986). Transition literature review on educational, employment and independent living outcomes. Champaign: The University of Illinois, The Transition Institute.

Harnisch, D. L., Fisher, A. T., Kacmarek, P. A., & DeStefano, L. (1987). Transition literature review: Educational, employment, and independent living outcomes (Vol 2). Champaign: The University of Illinois, The Transition Institute.

Heal, L. W., Colson, L. S., & Gross, J. C. (1984). A true experiment evaluating adult skill training for severely mentally retarded secondary students. American Journal of Mental Deficiency, 89, 146-155.

Keith, K. D., Schalock, R. L., & Hoffman, K. (1986). Quality of life: Measurement and programmatic implications. Nebraska City, NE: Region V Mental Retardation Services.

Kenny, D. A. (1975). A quasi-experimental approach to assessing treatment effects in nonequivalent control group designs. Psychological Bulletin, 82, 345-362.

Larson, R. C., & Kaplan, E. H. (1981). Decision-oriented approaches to program evaluation. In R. J. Woolridge (Ed.), Evaluating complex systems. New directions for program evaluation, 10 (pp. 49-68). San Francisco: Jossey-Bass.

Lipsey, M. W., Crosse, S., Dunkle, J., Pollard, J., & Stobart, G. (1985). Evaluation: The state of the art and the sorry state of the science. In D. S. Cordray (Ed.), Utilizing prior research in evaluation planning. New directions for program evaluation, 27 (pp. 7-28). San Francisco: Jossey-Bass.

Maltz, M. D., Gordon, A. C., McDowall, D., & McCleary, R. (1980). An artifact in pre-post designs: How it can mistakenly make delinquency programs look effective. Evaluation Review, 4, 216-225.

Office of Educational Research and Improvement, U.S. Department of Education, Center for Statistics. (1986, April). High School and beyond 1980 Sophomore cohort second follow-up (1984) data file user's manual. Washington, DC: National Center for Educational Statistics.

Salend, S. J., & Fradd, S. (1986). Nationwide availability of services for limited English-proficient disabled students. Journal of Special Education, 20, 127-135.

Schalock, R. L. (1986). Defining and measuring the quality of work and outside life. Paper presented at the Annual Conference of TASH, San Francisco.

Scheirer, M. A., & Rezmovic, E. L. (1983). Measuring the degree of program implementation: A methodological review. Evaluation Review, 8, 747-776.

Shadish, W. R., Jr., & Reis, J. (1984). A review of studies of the effectiveness of programs to improve pregnancy outcomes. Evaluation Review, 8, 747-776.